

# The STAR METRICS Project: Current and Future Uses for S&E Workforce Data

---

*Julia Lane*  
*National Science Foundation*

*Stefano Bertuzzi*  
*National Institutes of Health*

**DISCLAIMER**

The views expressed in this paper are the views of the authors and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

Documenting the impact of science investments requires documenting the activities of the science and engineering workforce. In most policy areas, such as health, labor or education policy, data typically provide some connection between investment decisions and outcomes at the appropriate behavioral unit of analysis. That is not the case in science policy. There are several major data challenges before science policy research achieves the same level of sophistication as these other fields. Most obviously, the information within administrative systems must be reoriented to connect investments with outcomes, and there must be broader and deeper collection of information on both inputs and outputs, particularly on the scientific workforce.

The STAR METRICS<sup>1</sup> program is focused on reorienting information within administrative systems to connect investments with outcomes. As such it is developing a data infrastructure that provides broader and deeper information on both inputs and outputs, recognizing that the appropriate unit of analysis is the scientific workforce. This paper discusses the current and future potential for the STAR METRICS program to collect data on the workforce.

## Background

The lack of data in science policy has not gone unnoticed. Indeed, OMB and OSTP have asked federal agencies to “develop outcome-oriented goals for their science and technology activities, establish procedures and timelines for evaluating the performance of these activities, and target investments toward high-performing programs. Agencies have been told to develop “science of science policy” tools that can improve management of their research and development portfolios and better assess the impact of their science and technology investments, and “In order to facilitate these efforts, Federal agencies, in cooperation with the Office of Science and Technology Policy and the Office of Management and Budget, should develop datasets to better document Federal science and technology investments and to make these data open to the public in accessible, useful formats”<sup>2</sup>

Science policy practitioners have been aware for some time that the current data infrastructure is inadequate. The National Science and Technology Committee’s Interagency working group on the Science of Science Policy identified lack of data as a critical gap in the implementation of evidence based science policy. In a workshop designed to roll out the Roadmap for federal investments, ten options were identified for measuring and tracking federal funding, there were three that participants overwhelmingly agreed upon. Nearly 94% of the participants were in favor of establishing a universal portal for dataset sharing (federal and non federal) that captures information about federal funding. Ninety-two percent agreed that a shared research environment, that would allow for data sets capturing information about federal funding to be integrated and analyzed by researchers, was a high priority. Eighty-nine percent of the participants agreed that federal funding agencies should

---

<sup>1</sup> Science and Technology in America’s Reinvestment – Measuring the EffectTs of Research on Innovation, Competitiveness and Science

<sup>2</sup> M-09-27 Memorandum for the Heads of Executive Departments and Agencies, August 4, 2009.

standardize their administrative records systems for initial awards as well as annual and final reports. These three options were also ranked as the Number One priority by the greatest percentage of participants.

The lack of data on the impact of science investments became even more clear to decision-makers with the passage of the 2009 American Recovery and Reinvestment Act (Goldston, 2009). Most of the estimates that were used for estimating the impact of science investments came from the Bureau of Economic Analysis' RIMS II model, which is derived from a 1992 input-output model of spending flows.. This approach functionally equates the impact of science to the impact of building a football stadium or an airport: the impact is derived from the demand side, and depends on the amount of spending on bricks and mortar and workers (Lane, 2009).

There are several reasons why we rely on such outdated models. The first is that U.S. scientific data infrastructure is oriented towards program administration rather than empirical analysis. The result is that 17 science agencies have 17 different data silos, with different identifiers, different reporting structures, and different sets of metrics. The second is that the focus of data collection is on awards, which are not the appropriate unit of behavioral analysis. Awards are the intervention of interest; it is the activities of the scientists that receive the awards that need to be followed. A third reason is that the current data infrastructure does not allow science investments to be coupled with scientific and economic outcomes. In particular, Grants.gov provides a unified portal to find and apply for federal government grants, but goes no further. Research.gov and science.gov provide information about research and development results associated with specific grants, and a consortium of federal agencies provides R&D summaries ([www.osti.gov/fedrnd](http://www.osti.gov/fedrnd)). Another obvious challenge is the fact that the reporting system is manual (with obvious quality implications) and relies on PIs to make reports during the active period of the award – despite the fact that the impact of science investments often results many years after the award has ended. Finally, despite the fact that science agencies believe that their impact includes both workforce and social impacts, there is no systematic tracking of the students supported by federal funds. A previous effort to collect R&D information on federal awards, RADIUS<sup>3</sup>, was discontinued in 2006.

It may be useful to illustrate the importance of linking output measures to the characteristics of scientists with a simple example. Science policy often uses a variety of metrics to determine country performance – one metrics is the measure of scientific publications per population. Suppose that measure is “low” for a given country relative to other countries. What is a policy maker to do in response to such information? Linked data enable more insights to be provided about the reasons for particular outcome – with different policy implications for each reason. For example, the data might show

1. The number is low because the country is making relatively heavy investments in computer science – and computer science research shows up in proceedings, not publications. The microdata links show that, conditional on the disciplines of the country's research community, the publishing rate is comparable to other countries. No policy action required.

---

<sup>3</sup> The Rand Database for Research and Development in the US

2. The number is low because the country has invested heavily in junior researchers, and it takes time for them to publish. The microdata links show that, conditional on the age of the country's research community, the publishing rate is comparable to other countries – and an expected outcome of investing in the future.

3. The number is low because the country has a number of old researchers at the end of their careers – and they've stopped publishing. The microdata links show that, conditional on the age of the country's research community, the publishing rate is comparable to other countries – but that other countries have many younger researchers. The policy implication might be to invest heavily in graduate fellowships and doctoral dissertations.

4. The number is low because a few highly published and internationally recognized researchers have left the country to go to another country. The microdata links show that the earnings for similarly qualified researchers in the country are lower than in the host country than in the other country. The policy implications might be to target salaries in that particular discipline.

In sum, without linked data, the examination of output metrics leads to pure speculation – and possibly incorrect policy decisions.

## **A Brief Overview of the STAR METRICS program**

The STAR METRICS program “Science and Technology for America’s Reinvestment: Measuring the EffecTs of Research on Innovation, Competitiveness and Science” (STAR METRICS) is led by an interagency Consortium consisting of NIH, NSF and OSTP. The goal of the program is to create a data infrastructure that will permit the analysis of the impact of science investments using administrative records as well as other electronic sources of data.<sup>4</sup> The new STAR METRICS program that is being developed by NIH and NSF under the auspices of OSTP is building a more scientific framework based on three principles.

The first is to use the right unit of analysis. Although current federal agency systems are built to administer awards, the new reporting demands on agencies require a new management information structure needs to be built with a different conceptual basis. The appropriate unit of analysis in that structure is scientists and clusters of scientists; the appropriate outcomes of scientific investments are the creation, dissemination and adoption of knowledge. The second is to use current technology. Fundamental transformations in digital technology can be used simultaneously reduce the need for manual reporting and facilitate the capture of appropriate outcomes – substantially improving the quality and reliability of the data infrastructure. The third is to collaborate with the scientific community. Domain scientists have the deepest

---

<sup>4</sup> Previously, NIH and NSF instituted a feasibility pilot program called simply STAR, at 7 volunteer universities that received ARRA funding. The pilot demonstrated that institutional and agency administrative data on jobs could be easily gathered and matched in a database. The pilot also demonstrated that web scraping could be utilized as a tool to track publications and citations that help show the long-term effects of Federal research grant funding

understanding of the appropriate data and metrics that should be used to describe the creation, transmission and adoption of knowledge in their fields. Social and behavioral scientists have the best understanding of how to theoretically and empirically tease out the impact of interventions.

The STAR METRICS project consists of two implementation phases:

- Phase I: Develop uniform, auditable and standardized measures of the impact of science spending (ARRA and non-ARRA) on job creation, using data from research institutions' existing database records.
- Phase II: Develop measures of the impact of federal science investment on scientific knowledge (using metrics such as publications and citations), social outcomes (e.g. health outcomes measures and environmental impact factors), workforce outcomes (e.g. student mobility and employment), and economic growth (e.g. tracing patents, new company start-ups and other measures).

## **STAR METRICS and Current Workforce Data**

In practical terms, Phase I of the STAR METRICS framework identifies how many scientists, including graduate students, undergraduate students and research staff, are supported by federal science funding. It also capture information about the jobs created from subawards, subcontracts and overhead. The balkanized nature of science funding has meant that this information has not existed in the past (Reedy, Litan, & Teitelbaum, 2001). STAR METRICS, building on its second and third foundational principle, works with collaborating research institutions and the Federal Demonstration Partnership to capture that information electronically (without personal identifiers).

The process by which these data are electronically captured is illustrated in Figure 1. The basic idea is simply to track the financial traces of award activity through the administrative systems of each institution. Since all expenditures associated with an award are tracked by means of an account code, including HR charges, institutions can produce the fourteen key data elements identified in Figure 2 with minimal burden.

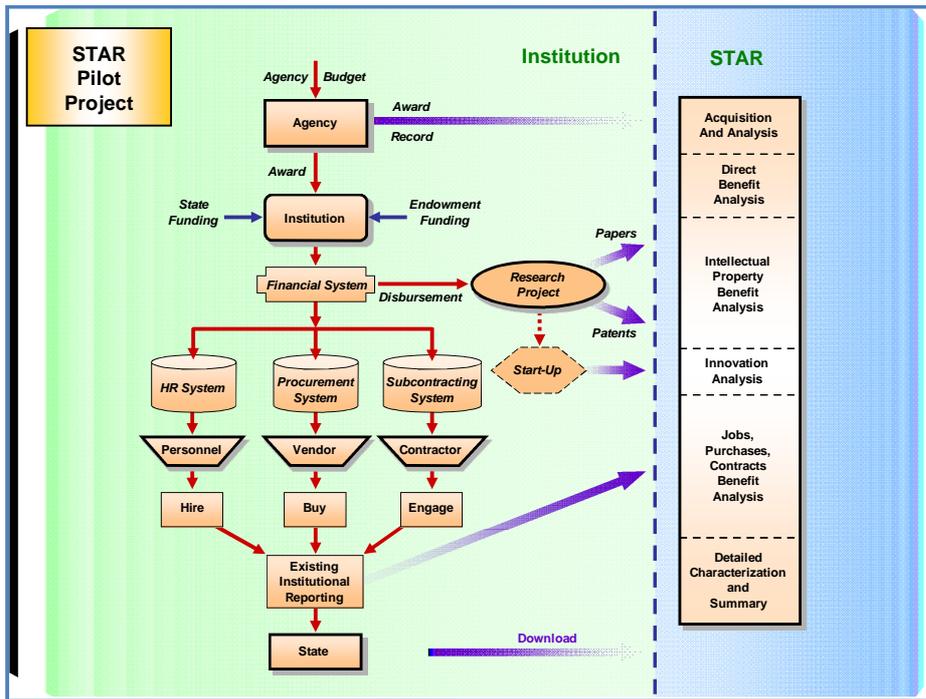


Figure 1

### Star Metrics Phase 1 – 14 Requested Data Elements

\*Each file has a period start and end date that reflects the quarter or time period in which the transaction occurs. NOT the start and end date of the award being reported.

Description	Element ID	Item
Information on Scientists and Awards	1	Unique Award#
	2	Recipient Account#
	3	Overhead charged
	4	De-identified Employee ID#
	5	Occupational Classification
	6	Proportion of earnings allocated to award
	7	FTE status
Subcontracts and subawards	1	Unique Award#
	2	Recipient Account#
	8	Sub Award Recipient Duns #
	9	Sub Award Payment Amount
Payments to Vendors	1	Unique Award#
	2	Recipient Account#
	10	Vendor Duns #
	11	Vendor Payment Amount
Each file has a period start and end date that reflects the quarter or time period in which the transaction occurs	12	Period Start Date
	13	Period End Date
Information on Overhead	14	Proportion of overhead associated with salaries (from indirect cost rate proposal)

Figure 2

Those data elements are all that is needed to create measures of employment by occupation over time. The conceptual framework in Phase I, which is described in more detail at

<https://www.starmetrics.nih.gov>, draws heavily on the LEHD Program at the US Census Bureau, which integrates Federal administrative and survey data and participating state Unemployment Insurance (UI) wage records and Quarterly Census of Employment and Wages (QCEW) data. The integration of these micro-data results in a longitudinal database of workers and firms, not only comprising a time series of information on individual workers and firms, but also tracking the movement of workers across firms. That program also started as a small pilot program (Lane, Burgess, & Theeuwes, 1997) and then later grew into a national program (Abowd, Haltiwanger, & Lane, 2004).

The STAR METRICS team and its collaborators use these data to generate tables, graphs and maps of the jobs directly supported by science funding (See Figures 3 and 4). Over 60 research institutions have signed participation agreements with the STAR METRICS program, and reports are being generated on a quarterly basis for those institutions. Many more institutions are in the process of joining.

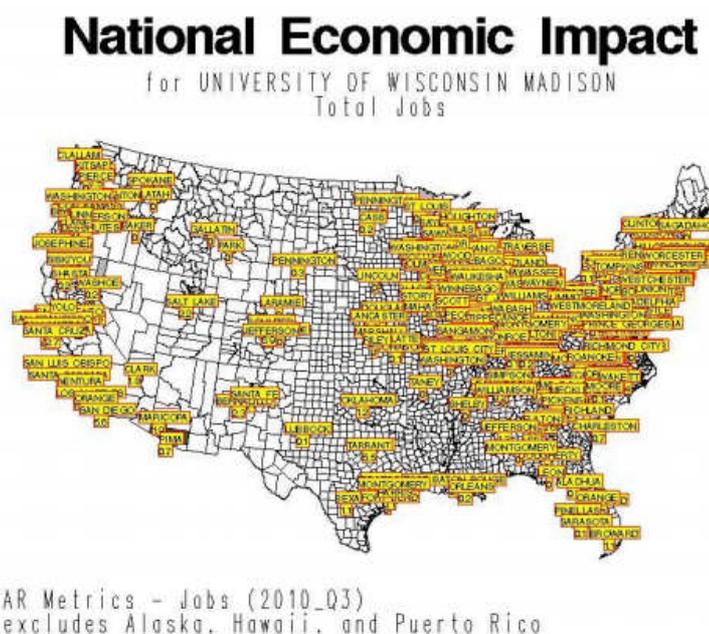


Figure 3

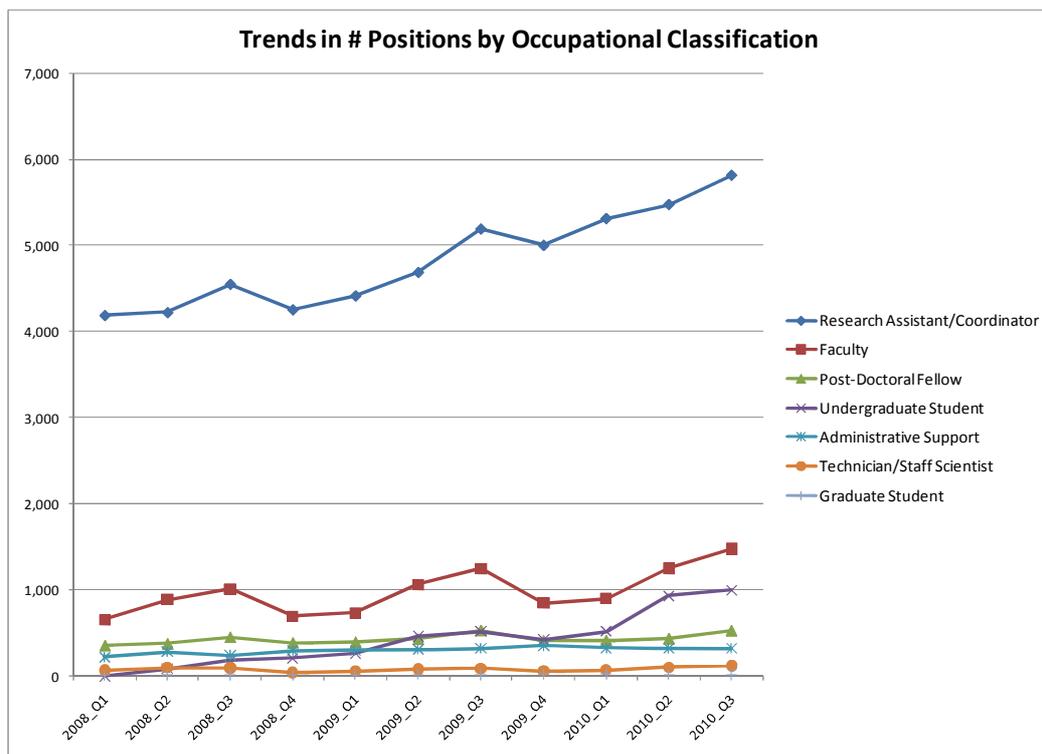


Figure 4

## STAR METRICS and Potential Workforce Data

It is, of course, critical to go beyond the counting of the science and engineering workforce described in Phase I. It is also of interest to describe their scientific activities, their mobility, and their employment and earnings trajectories. The next phase, Phase II of STAR METRICS, is intended to do just that. It aims to leverage the fundamental transformation in digital technology to capture the scientific, social, economic, and workforce impacts – and use scientific advances in visual analytics to convey them in intuitive ways. The complex nature of the scientific enterprise means that Phase II is likely to take many years.

The consultation with the scientific community about possible data elements and sources only began in Fall 2010, beginning with a meeting of Vice Presidents for Research of interested institutions on October 22 at the National Press Club.<sup>5</sup> However, several possible initial steps have been proposed and met with some enthusiasm by participating universities.

<sup>5</sup> The list of attending institutions is provided in the Appendix.

One approach is to use existing administrative data, such as the US Patent Office data, to link patent data and the associated critical publications to their intellectual provenance in federally funded research. (Fleming & Torvik, 2009). That research, which also links patents to the patent assignees and the technology class of the research, can be used to identify which patents belong to which inventor, and hence to trace the flow of knowledge, trace the mobility of PIs to the private sector, and identify breakthroughs (both patents and inventors). For example, the research has permitted the identification of the assignees of patents attributable to the work of NSF and NIH funded PIs at the University of Pennsylvania (Figure 5).

739 inventors							
	104	AMGEN INC					
	58	UNIVERSITY OF PENNSYLVANIA					
	57	GILEAD SCIENCES INC					
	34	VERENIUM CORPORATION					
	23	CARGILL INCORPORATED					
	23	KOSAN BIOSCIENCES INC					
	19	SUGEN INC					
	19	WYETH					
	17	SARNOFF CORPORATION					
	16	AVENTIS PHARMACEUTICALS INC					
	14	THE TIMKEN COMPANY					
	14	THE UNITED STATES OF AMERICA AS REPRESENTED BY THE SECRETARY OF THE NAVY					
	12	MIGENIX CORP					
	12	ABBOTT LABORATORIES					
	12	THERAVANCE INC					

Figure 5

Another possibility is to match the administrative records of universities with the data of statistical agencies (such as the Statistics of Income Division or the Census Bureau)<sup>6</sup>. Data on graduate and undergraduate students as well as postdoctoral students could be matched to the firms for which they work subsequent to their federal support. It would then be straightforward to generate their employment and earnings trajectories, as well as document the firms and industries in which they work, and the competitiveness and survival of those firms.

Other possibilities abound. Academic researchers, particularly those funded by SciSIP, have collected large bodies of data on such scientific and innovation outcomes as clickstream data, citations, patents, patent applications, business startups and IPOs. In some case, such as the Zucker/Darby project<sup>7</sup>, those data have been successfully linked to outcome measures. And

<sup>6</sup> Any matches would require the informed consent of the individual involved.

<sup>7</sup> The Zucker/Darby project integrates data on government grants, journal articles, dissertations, patents, venture capital, initial public offerings, and other firm data. It links to major public databases via widely used financial market identifiers. It links the data to Census firm and worker databases by a concordance for use by researchers with access to the Census data. The database will have three tiers: a public graphics-based site primarily oriented toward policymakers and the media, a public site providing access to researchers for downloads and database queries limited to the public constituent databases or aggregates derived from the licensed commercial databases,

existing cyberinfrastructure can create flow reports of citations, patents, and publications from webscraping that could be used for the automated reporting of PI activities both during and well beyond the period of their grant, as well as be used for the automated generation of biosketches.

There is extraordinary interest in the scientific community in such research. There are some 200 SciSIP PIs who can be tapped to participate. Faculty participating in the FDP program have volunteered to become engaged. Science agencies across the world are emulating the approach. It is certainly feasible to provide researcher access to such confidential micro data as might be generated by Phase II of the STAR METRICS program. The Institute for Quantitative Social Science at Harvard is one example. The San Diego Super Computer Center is another. NSF's Science Resources Statistics Division has contracted with the NORC/University of Chicago's Data Enclave to provide researcher access to microdata. The Statistics of Income Division at IRS is investigating the same access modality. In general, it will be necessary to build an open access, cyberinfrastructure enabled, collaborative environment which can be used so that the research community can collaborate with the federal agencies to generate summary indicators about where science investments have been and are being made, together with information about the economic, social and scientific impacts over space and time. Summary

## Summary

It has become critical to develop an evidence basis for science policy. From a practical point of view, science agencies have a looming imperative to document the impact of the nearly \$20 billion in R&D investments embodied in the 2009 American Recovery and Reinvestment Act (ARRA). It is also clear that the Federal budget environment is likely to be extremely competitive for the foreseeable futures. In order for a case to be made that investments in science have value relative to investments in education, health or the workforce, an analytical and empirical link has to be made between those investments and policy-relevant outcomes. The STAR METRICS program is intended to be a collaboration between science agencies and research institutions to do just that.

[Institutional Attendees: Oct 22 Workshop](#)

Arizona State University

Boston University

Brown University

California Institute of Technology

Case Western Reserve University

College of Charleston

Colorado State University

Dartmouth College

Emory University

George Mason University

Georgia Institute of Technology

Harvard

John Hopkins University

Michigan State University

Massachusetts Institute of Technology

Northern Arizona University

Northwestern

New York University

Oregon Health & Science University

University of Pennsylvania

Purdue University

Rockefeller University

State University of New York, Stony Brook

Syracuse University

Temple University

Texas A and M University

The Research Foundation, State University of New York

University of Alabama

University of Arizona

University of California, San Diego

University of Chicago

University of Delaware

University of Illinois, Urbana Champaign

University of Kansas

University of Minnesota

University of Minnesota

University of Missouri, Columbia

University of Southern California

University of Texas, Austin

University of Texas, San Antonio

University of Virginia

Vanderbilt University

Oakridge National Lab

## References

- Abowd, J., Haltiwanger, J., & Lane, J. 2004. Integrated Longitudinal Employer-Employee Data for the United States. *American Economic Review*, 94(2): 224-229.
- Fleming, L., & Torvik, V. 2009. From Grant to Commercialization: an integrated demonstration database which permits tracing, assessing and measuring the impact of science funding: National Science Foundation.
- Goldston, D. 2009. Mean what you say, *Nature*.
- Lane, J. 2009. Assessing the Impact of Science Funding. *Science*, 324(5932): 1273-1275.
- Lane, J., Burgess, S., & Theeuwes, J. 1997. *The Uses of Longitudinal Matched Employer/Employee Data in Labor Market Analysis*. Paper presented at the American Statistical Association, Anaheim, CA.
- Reedy, E. J., Litan, B., & Teitelbaum, M. 2001. The Current State of Data on the Science and Engineering Workforce, Entrepreneurship, and Innovation in the United States. In K. H. Fealing, J. Lane, J. H. Marburger, & S. Shipp (Eds.), *The Handbook of Science of Science Policy*: Stanford University Press.